



Supplementary Specification for Ingest

ECHO 10.0

Document Version 10.10

**Global Science and Technology, Inc
7855 Walker Drive, Suite 200
Greenbelt, MD 20770 USA
Phone (301) 474-9696
Fax (301) 474-5970**

REVISION HISTORY

Date	Version	Brief Description of Change	Author
05/16/07	10.1	Initial version for ECHO 10, Iteration 3	Judith Wright
05/21/07	10.2	Incorporate comments from Use Case/Requirements Review	Judith Wright
05/29/07	10.3	Incorporate initial comments from ETC (Beth Weinstein)	Judith Wright
06/06/07	10.4	Updating for Iteration 3.	Jason Gilman
06/11/07	10.5	Integrated changes from the BMGT requirements review	Michael Pilone Judith Wright
07/17/07	10.6	Updated for ECHO 10, Iteration 4	Judith Wright Dan Pilone Mike Pilone Jason Gilman Lisa Pann
07/31/07	10.7	Incorporate external review comments	Judith Wright
09/03/07	10.8	Updated for ECHO 10, Iteration 5, including, but not limited to: Browse, Ingest Reporting, Audit Trail	Judith Wright
9/18/07	10.9	Added fix for NCR 10003048 to section 4.4.1.f	Lorena Marsans
9/19/07	10.10	Incorporate ETC comments	Judith Wright

TABLE OF CONTENTS

1. GLOSSARY	1
2. ECHO ACTORS.....	1
2.1. Actor Descriptions	1
2.2. Actor Hierarchy.....	3
3. ACTION DETAILS	5
3.1. Error and Exception Handling	5
3.1.1 Ingest Logging.....	5
3.2. Audit Trails	5
3.2.1 General	5
3.2.2 Collection, Granule, Browse Deletes.....	6
3.3. Notifications.....	7
3.4. Ingest Reporting.....	7
3.4.1 Ingest Detail Report.....	7
4. BUSINESS RULES.....	7
4.1. File and Job Handling.....	7
4.1.1 Starting/Stopping Ingest	7
4.1.2 Ingest Configuration	8
4.1.3 File Handling.....	8
4.1.4 Directory Structure.....	10
4.2. Establishing a Provider.....	11
4.3. General Metadata	11
4.3.1 Spatial	12
4.3.2 Temporal	13
4.4. Collections.....	13
4.4.1 General Collections	13
4.4.2 Temporal	14
4.4.3 Spatial	14
4.4.4 Platforms	14
4.4.5 Platform/Instrument/Sensor Characteristics.....	14
4.4.6 Instruments.....	14
4.4.7 Sensors	14
4.4.8 Additional Attributes.....	15
4.4.9 Browse Reference.....	15
4.5. Granules.....	15

4.5.1	General Granules	15
4.5.2	Spatial	16
4.5.3	Temporal	16
4.5.4	Platform/Instruments/Sensors	16
4.5.5	Additional Attribute Value	16
4.5.6	Browse Reference	16
4.6.	Browse	16
4.7.	Actions	17
4.7.1	General Item	17
4.7.2	Collections	18
4.7.3	Granules	19
4.7.4	Browse Images	20
4.8.	Ingest Reporting	21
4.8.1	Ingest Detail Report	21
4.9.	Provider-Specific Adapters	22
5.	INGEST DATA ENTITIES	23
6.	INGEST JOB STATES	23
7.	EMAIL NOTIFICATIONS CONTENT	23
7.1.	Operational Emails	24
7.1.1	Start Ingest	24
7.1.2	Stop Ingest	24
7.2.	Ingest Job-Specific Emails (per Data Provider)	24
7.2.1	No Input Received	24
7.2.2	Ingest Job Started	25
7.2.3	Ingest Job Completed	25
7.2.4	Out of Sequence Job	26
7.3.	Ingest Errors/Failures	27
7.3.1	Error in Input Package and/or Table of Contents	27
7.3.2	Fatal Error	27
7.3.3	Error in Manifest File	27
7.3.4	Job Requires Sequencing and Wait Time Exceeded	28
7.4.	Pause/Resume Notifications	28
7.4.1	Input Detection Paused for Provider	28
7.4.2	Input Detection Resumed for Provider	29
7.4.3	Provider Paused	29
7.4.4	Provider Resumed	30
7.5.	Report Notifications	30

7.5.1 Ingest Detail Report.....	30
8. SECURITY	30
9. VALIDS	31
10. DESIGN CONSTRAINTS.....	31
11. NOTES	31
12. ISSUES	31
12.1. General Issues	31
12.1.1 File Handling.....	32

1. GLOSSARY

- a. Authorized Provider – A provider on record that has been approved by Operations.
- b. Client Partner – Client application, a program that uses Echo.
- c. Types of Ingest Jobs:
 - 1) Pending—jobs that Ingest has detected but has not yet received all the information necessary for ingesting metadata
 - 2) Ready—jobs for which all necessary information has been received and are ready for ingesting as soon as a pipeline opens up
 - 3) Executing—jobs that are in the process of ingesting metadata
- d. Thresholds
 - 1) Quiet Period

The amount of time that a file must have no modifications and be complete, before it is added to the Job. E.g., file FOO.XML received from GSFC must not have been modified for the past X units before it is added to the Job. ("Units" are units of time). This threshold is configurable by the Ingest Operator on a per-provider basis.
 - 2) No Input Threshold

The maximum amount of time that can elapse without receipt of Ingest input before the Ingest Operator is notified that no input has been received recently. This threshold is configurable by the Ingest Operator on a per-provider basis.
 - 3) Resource Wait Threshold

The maximum amount of time that can elapse without receipt of resources for a file. (i.e., browse images)
 - 4) Out of Sequence Job Threshold

The maximum amount of time that ingest will wait if receiving a job out of sequence for the other jobs in sequence before notifying the operator.

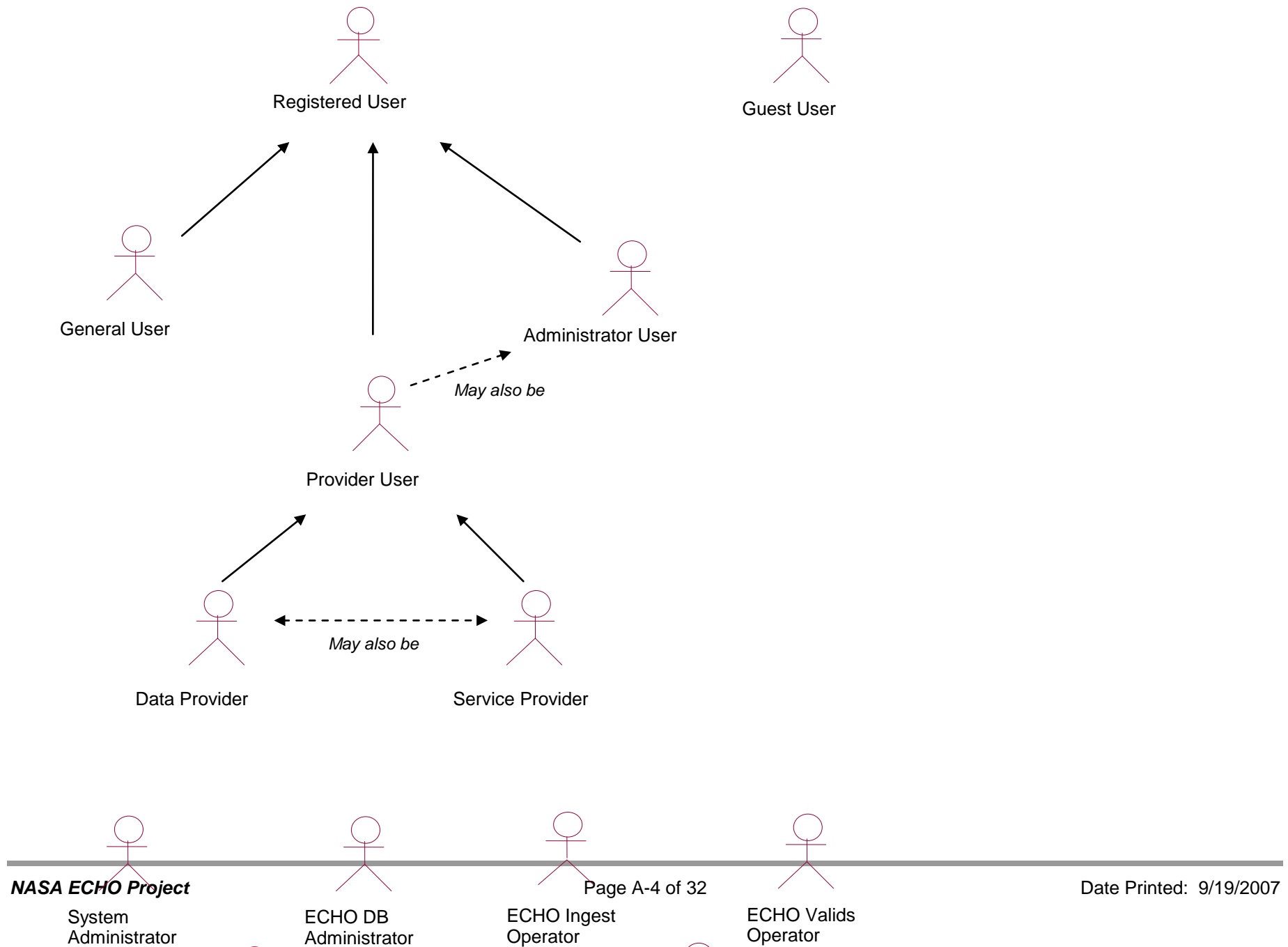
2. ECHO ACTORS

2.1. Actor Descriptions

- a. Guest User—A user who does not login to use ECHO
- b. Registered User—A user that has been registered in ECHO, has an ECHO user ID assigned which is used to login to ECHO
 - 1) General User—Registered User who is neither a Provider User nor an Administrator User

- 2) Provider User—Registered User of ECHO representing a data and/or service provider who has been given privileges to Provider transactions
 - a) Service Provider—Registers their services in the service registry
 - b) Data Provider—Puts metadata into the registry
- 3) ECHO Administrator—Registered User who has super user privileges in ECHO, approves accounts, etc.
- 4) *Note: A Registered User may be both a Provider User and an Administrator.*
- c. System Administrator—Starts, stops, and configures echo deployment
- d. ECHO DB Administrator—Migrates data, creates initial databases, backs up DBs, etc.
- e. ECHO Ingest Operator—Responsible for ingesting data into ECHO
- f. ECHO Validates Operator—Responsible for transforming ECHO holdings to V0 Validates
- g. ECHO Health Operator—Monitors health and security logs
- h. Provider Configuration Monitor—Handles provider requests, applications, etc.
- i. Extended Services Monitor—Handles service interface submissions, implementations, etc.

2.2. Actor Hierarchy



3. ACTION DETAILS

This describes in more detail actions that can occur within ECHO Ingest

3.1. Error and Exception Handling

With the exception of metadata item validation failures, all errors are logged in the Ingest Log.

3.1.1 Ingest Logging

- a. Logging is done to text log files.
- b. There will be three types of log files.
 - 1) A job log file will contain log entries relevant for a specific job.
 - a) There will be one log file for each job.
 - 2) A provider log file will contain log entries relevant for a specific provider.
 - a) There will be one log file for each provider.
 - b) The provider log file will contain all of the job log entries for that provider.
 - 3) A root log file will contain global ingest events.
 - a) The root log file will not contain entries from the provider or job log files.
- c. All the log entries will include the following "Standard Logging Fields":
 - 1) Timestamp
 - 2) Component logging the information
 - 3) Error code if appropriate
 - 4) Human readable message describing the log entry.
- d. Depending on the nature of the event, the data collected may be limited to the *Standard Logging Fields* or may include identifying information about the data entity impacted by the event and detailed information about data changes that occurred as result of the event.

3.2. Audit Trails

3.2.1 General

- a. Ingest shall store the Audit Trail information in the database
- b. Ingest shall keep the Audit Trail information indefinitely

- c. No database constraint or other data validation checking should be done on data moved/copied to audit trail

3.2.2 Collection, Granule, Browse Deletes

- a. Entries in the Audit Trail will be made for the item in the delete request only. Deletes resulting from cascades will not be recorded in the Audit Trail. E.g., if a request was submitted to delete a collection that owns one or more granules, that collection deletion will be recorded in the Audit Trail but the granules that are deleted as a result of cascading will not be recorded.
- b. Ingest shall record the following information in the audit trail for collections:
 - 1) Short Name
 - 2) Version ID
 - 3) Long Name
 - 4) Dataset ID
 - 5) Dates/Times the Data Providers send:
 - a) Insert Time
 - b) Update Time
 - c) Delete Time
 - 6) Local Dates/Times
 - a) Insert Time
 - b) Update Time
 - c) Delete Time
- c. Ingest shall record the following information in the audit trail for granules:
 - 1) Granule UR
 - 2) Short Name
 - 3) Version ID
 - 4) Long Name
 - 5) Dataset ID
 - 6) Dates/Times the Data Providers send:
 - a) Insert Time
 - b) Update Time
 - c) Delete Time

- 7) Local Dates/Times
 - a) Insert Time
 - b) Update Time
 - c) Delete Time

3.3. Notifications

- a. Notifications from Ingest are done using email.
- b. All notifications are logged.
- c. Provider notifications are sent to the Provider configured contacts.

3.4. Ingest Reporting

3.4.1 Ingest Detail Report

At the end of each Ingest Job, an Ingest Detail Report is automatically generated and stored at both an ftp site and in the Archive to be later retrieved by the Ingest Operator. Reports are stored in the Archive for an indefinite amount of time based on ECHO Operations policy. The Archive will reside in the Job Directory.

The content of this report is defined in the Business Rules.

4. BUSINESS RULES

4.1. File and Job Handling

4.1.1 Starting/Stopping Ingest

- a. Ingest shall only allow a single ingest process to run against the same configuration.
- b. Only one job from a provider can be loaded at a time
- c. Ingest shall validate the configuration prior to starting the Ingest process.
- d. Ingest shall notify the operator via email on Ingest startup.
- e. Ingest shall support graceful shutdown during processing and maintaining its state and later resuming activities.
 - 1) Shutting down shall occur upon completion of processing of the current metadata segment in each pipeline.

- 2) The states that will be maintained are:
 - a) Job hold states
 - b) Job activity state
 - c) Processed files
 - d) Provider state
- f. Ingest shall notify the operator via email of Ingest shutdown.

4.1.2 Ingest Configuration

- a. The configuration file will be a plain text XML file.
- b. Ingest shall support changing the configuration while shutdown.
 - 1) Ingest shall support editing the configuration file only when shutdown.
- c. Changes made to the configuration file while Ingest is running may be lost.
- d. Ingest shall apply configuration changes to jobs after restarting
- e. See Ingest Configuration Entities

4.1.3 File Handling

- a. Ingest shall support filenames up to 255 characters long.
- b. Ingest shall process files in the following order:
 - 1) Browse inserts/replacements
 - 2) Collection inserts/replacements
 - 3) Collection partial adds
 - 4) Collection partial updates
 - 5) Collection partial deletes
 - 6) Collection deletes
 - 7) Granule inserts/replacement
 - 8) Granule partial adds
 - 9) Granule partial updates
 - 10) Granule partial deletes
 - 11) Granule deletes

-
- 12) Browse deletes
 - c. Data Provider Notification of Ingest Job Start
 - 1) The Data Provider will be notified via email of their ingest jobs starting or stopping if they have requested the emails be sent.
 - a) This request will be recorded in the Data Provider's Configuration File and is on a per-Provider basis, not per-job basis
 - b) The Ingest Operator will update the Data Provider's Configuration File with this information based on the Data Provider's preference.
 - 2) Ingest will notify via email the Data Provider when one of their jobs leaves the SEQUENCE_WAITING state.
 - 3) The notification email will be sent to the Data Provider's configured email addresses (in the Provider's Configuration File)
 - 4) Ingest shall provide a way for the Data Provider to specify it s/he wishes to be notified via email when one of his/her ingest jobs has started
 - 5) If a Data Provider has requested it, Ingest shall notify him/her via email when one of his/her ingest jobs has started
 - d. Ingest shall indefinitely hold a job as pending if it is out of sequence.
 - 1) Ingest will continue to detect input and resume processing jobs when the correct sequence number has been detected
 - 2) Ingest shall notify the Ingest Operator and Data Provider if an out of sequence job has been pending for a configurable amount of time.
 - e. If a configurable amount of time has passed since the last file or package was detected from a Data Provider (No Input Threshold), then the Ingest Operator and Data Provider will be notified via email.
 - 1) This Threshold may be disabled on a per-provider basis.
 - 2) This Threshold shall be configurable on a per-provider basis.
 - f. Ingest shall demarcate transactions per individual metadata item.
 - g. Ingest shall continue processing metadata items in a file even if metadata items previously in that file were rejected.
 - h. Ingest shall halt a job if a fatal error is detected.
 - 1) A fatal error is defined as:
 - a) A configuration error
 - b) Programming error
 - c) Resource error (e.g., database unavailable or file system full)
 - 2) Ingest shall notify the Ingest Operator of the error if fatal error allows Ingest to notify the Ingest Operator
 - 3) Ingest shall attempt to shutdown gracefully if fatal error allows Ingest to shut it down gracefully as defined in section 4.1.1, Starting/Stopping Ingest, item e
 - i. Ingest shall allow deletion of Jobs that are not in the loading or later job states
 - j. Ingest must support the ECHO format, ECHO 8 DTD, and BMGT DTD.
-

- k. Ingest will validate input files:
 - 1) Check to see if it is a properly formed XML file
 - 2) Validate the input file with its corresponding DTD or schema
 - 3) If this validation fails, the input file will be rejected.
- l. When an input file is rejected, an error will be recorded and included in the Ingest Detail Report
- m. When a job has been completed, an Ingest Detail Report will be generated as defined in the *Ingest Detail Report section*

4.1.4 Directory Structure

Files will be stored in the following directory structure

- Workspace/
 - config.xml (Ingest configuration file)
 - ingest-instance.lck (Ingest lock file)
 - ingest.log
 - jobs/
 - <PROVIDER>
 - <PROVIDER>.log
 - <YYYY>
 - ↳ <MM>
 - <DD>
 - job-<GUID>
 - » processing/
 - <Granule1>.xml
 - ...
 - » resources/
 - <Browse1.hdf>
 - <Browse2.hdf>
 - ...
 - » job.log (Job log file)

-
- » Job report(s)
 - » <Input metadata file(s)>
- a. There will be a single workspace directory for any ingest configuration.
 - 1) The Ingest Configuration file will be contained at the top level of the workspace directory
 - 2) Ingest will create a lock file to lock the configuration and prevent multiple ingest processes from running against the same configuration.
 - 3) The main ingest log file will be kept in the workspace directory.
 - b. The Jobs directory will contain all of the files for an ingest job in separate directories for each provider.
 - 1) The provider directory will contain the provider log file.
 - c. The provider directories will contain the jobs in a hierarchy of directories for the year, month, and day the job was created on.
 - d. The Job directory will be named using the job GUID.
 - 1) Input metadata files for the job will be kept in the job directory.
 - 2) Processing files derived from the input metadata files will be kept in a Process directory.
 - 3) Resources referenced by the input metadata will be kept in a Resources directory.
 - 4) The job log will be kept in the job directory.
 - 5) Any generated reports for the job will be kept in the job directory.

4.2. Establishing a Provider

- a. Provider-specific information shall be configured per provider
- b. FTP directories shall be configured for and used by providers for delivering their metadata files
- c. ECHO defaults for latitude and longitude:
 - 1) Latitude range: -90 to 90 degrees
 - 2) Longitude range: -180 to 180 degrees
 - 3) If not specified, ECHO defaults the latitude and longitude resolution to 0.0001
 - a) In Cartesian, the resolution unit is degrees
 - b) In Geodetic, the resolution unit is meters

4.3. General Metadata

- a. All date/time fields shall be stored in GMT

- b. Required fields shall be enforced as defined by the metadata schema
- c. Ingest shall reject metadata items individually.
- d. The ECHO system does not do a unification or mapping of the additional attributes' names among the Data Partners.
- e. Collection metadata must be in the ECHO database before any of that granule's metadata associated with the collection is accepted in ECHO.

4.3.1 Spatial

- a. Ingest will support multiple spatial types.
 - 1) Geometry
 - 2) Global
 - 3) Orbit
- b. Spatial geometry is validated according to Oracle's Spatial User's Guide and Reference
- c. Each provider is allowed to have different spatial representation; however, each collection may define a single spatial representation for its granules
- d. ECHO currently supports two coordinate systems for geometric data
 - 1) Cartesian
 - 2) Geodetic
- e. Each data provider must use only one coordinate system when constructing the spatial area coverage for a collection or granule
- f. To avoid misinterpretation of a data provider's spatial data, Ingest will not manipulate any of the spatial input data
- g. At this time, a spatial search is not allowed on a collection or granule that uses a circle spatial coverage representation.
- h. In the Cartesian system, points are connected in a straight line
- i. In the Geodetic system, points are connected in an arc using the shortest distance between the points.
- j. Point: ECHO will receive, store, and support the search on spatial data representing one or more points
- k. Line: ECHO will receive, store, and search spatial data representing one or more lines
- l. Polygon: ECHO will receive, store, and search spatial data representing a polygon, a polygon with hole
 - 1) Points must be specified in a clockwise direction
 - 2) Single Polygons (GPolygons):
 - a) A single polygon can only have one outer ring that represents the area surrounded within
 - b) A single polygon with a hole can only have one outer ring that represents the area surrounded within.
 - c) A single polygon with a hole can have multiple inner rings that represent holes.

- d) An inner ring must be completely contained by an outer ring.
- e) Twisted Polygons are invalid
- f) In the Cartesian system, the area may not cross the International Date Line or Poles.
- g) In the Geodetic system, the area may cross the International Date Line and/or Poles
- h) In the Geodetic coordinate system, the area is not allowed to have an area greater than or equal to half the area of the Earth.
- m. Bounding Box: ECHO will receive, store, and search spatial data representing a bounding box or multiple bounding boxes
 - 1) Bounding boxes are subject to the same constraints as Polygons
 - 2) Multiple bounding boxes are stored as multiple single polygons
- n. For any spatial data expressed using latitude/longitude, the range is expressed in degrees and the tolerance/resolution in meters.

4.3.2 Temporal

- a. For providers using the ECHO Format, Ingest requires date & times to be formatted according the XML Schema Standard dateTime format
- b. For providers using a legacy DTD:
 - 1) Ingest must support varying time formats; however only one format per provider will be supported
 - 2) The time zone of the temporal information must be GMT/UTC
 - 3) If the date/time format does not match the data provider's specified format, Ingest will reject the item

4.4. Collections

4.4.1 General Collections

- a. Short name and version ID combined must be unique within a provider
- b. Long name and version ID combined must be unique within a provider
- c. Dataset ID shall be a unique identifier within a provider. If a provider does not support a dataset ID, the dataset ID shall be constructed as follows:
 - 1) ECS Providers:
 - a) Long name and version ID shall be combined to produce a dataset ID
 - 2) Non-ECS Providers:
 - a) Short name and version ID shall be combined to produce a dataset ID
- d. Local last update date is set to the current time on any type of modification
- e. Provider update date shall be used to detect out of date records

- f. Collection ShortName shall be limited to uppercase characters, numbers and “_”.

4.4.2 Temporal

- a. Three types of temporal data for collections will be supported:
- 1) Single day time
 - 2) Range day time
 - 3) Periodic day time
- b. A collection may have only one temporal type, but may contain multiple instances of that type

4.4.3 Spatial

- a. For providers not using the ECHO Format and the coordinate system is not specified, the provider's default coordinate system will be used
- b. Granule spatial representation shall be one of the following:
- 1) Cartesian
 - 2) Geodetic
 - 3) Global
 - 4) Orbit
 - 5) NoSpatial

4.4.4 Platforms

- a. Platforms are uniquely identified by short name.

4.4.5 Platform/Instrument/Sensor Characteristics

- a. Characteristics are uniquely identified by name

4.4.6 Instruments

- a. Instruments are uniquely identified by short name

4.4.7 Sensors

- a. Sensors are uniquely identified by short name

4.4.8 Additional Attributes

- a. Ingest will define unique Additional Attributes at the collection level
 - 1) Additional attributes shall be uniquely identified by name
- b. Additional attributes cannot be removed if a child granule is referencing the attribute.
- c. *See ECHO 10 Supplementary Specification for more information on Additional Attributes and Valid*

4.4.9 Browse Reference

- a. Browse references are uniquely identified by provider browse ID.
- b. The browse image referenced must exist for the provider.
 - 1) The provider browse ID is given to reference a browse image.

4.5. Granules

4.5.1 General Granules

- a. Every granule must belong to a collection
 - 1) The collection referenced by the granule must be in the ECHO database.
- b. If not specified in the granule, the Restriction Flag field value shall default to the value in the parent collection
 - 1) If the restriction field is copied from the parent collection, the comment field value shall be copied from the parent collection
- c. If the restriction flag is not set for a granule and there is no collection default, then the restriction flag will be set to NULL
- d. If the Visibility Flag field value is not specified, it will default to the provider's default Visibility Flag.
- e. If the Data Format field is not specified in the metadata, the value will be copied from the parent collection
- f. A granule's campaigns must be a subset of its associated collection's campaigns.
- g. Campaign references shall be uniquely identified by Short Name at the granule level
- h. Online Access URLs shall be uniquely identified by the URL at the granule level
- i. Online Resources shall be uniquely identified by the URL at the granule level
- j. Measured Parameters shall be uniquely identified by Parameter Name at the granule level
- k. Two-dimensional coordinate system types are defined by ECHO operations and must be enabled for a collection before it may be used in granule ingest.

4.5.2 Spatial

- a. Granule spatial representation must match the granule spatial representation specified in the parent collection

4.5.3 Temporal

- a. Two types of temporal information represent the time when granule data were collected:
 - 1) Single day/time
 - 2) Range day/time
- b. A granule's temporal must be in the range of its collection's temporal

4.5.4 Platform/Instruments/Sensors

- a. A granule's platform/instrument/sensor configurations must be a subset of its associated collection's platform/instrument/sensor configurations.
- b. A granule's instrument operation modes must be a subset of its associated collection's instrument operation modes.
- c. A granule's platforms' characteristics are defined by its primary collection
- d. A granule may define instrument and sensor level characteristic value
 - 1) Characteristics' names must be a subset of its associated parent collection's instrument and sensor characteristics

4.5.5 Additional Attribute Value

- a. A granules' Additional Attributes must be a subset of their associated collection's Additional Attributes by Name.
- b. A granule's additional attribute value shall override the value of the parent collection

4.5.6 Browse Reference

- a. A granule may be associated with 0 to many browse files
- b. A browse file may be referenced by more than one granule.
- c. Browse references are uniquely identified by provider browse ID
- d. The browse image referenced must exist for the provider.
- e. The provider browse ID is given to reference a browse image.

4.6. Browse

- a. Browse image file names must be unique per provider.
- b. Browse images are uniquely identified by provider browse id

- c. The browse image file must be delivered with the browse image metadata.
- d. The size of the browse image file must match the FileSize given in the browse image metadata.
- e. The ECHO system will allocate the storage for browse files, build a browse image URL, and update the database to associate the browse URL to its item record.

4.7. Actions

4.7.1 General Item

- a. For any repetitive item identification in the metadata input, only one item bearing the most recent update date with its complete information will be ingested into the database. The rest of the items and their associated information will be ignored even if they contain different data information and will result in a warning in the ingest detail for that item.
- b. Insert**
 - 1) Item local insert date is set to the current time
 - 2) Item local last update date is set to the current time
 - 3) If the item to be inserted already exists, the insert will be handled as a "Replace"
- c. Replace**
 - 1) Item local last update date is set to the current time
 - 2) Item shall not be updated if the provider last update date/time of the stored item is after the new item's provider last update date/time
 - a) If the last update date/time of the stored item is before or the same as the new item's provider last update date/time, the item will be ignored and a warning will be generated
- d. Delete**
 - 1) The item shall be removed from the DB immediately
 - 2) The deleted items' identification and deletion date will be kept in the ECHO database for data history auditing purpose
 - 3) Deleted items cannot be restored
- e. Partial Update**
 - 1) Item local last update date/time is set to the current date/time
 - 2) Ingest will set the provider's last update date to the time in the update request
 - a) Ingest will support overriding this behavior via a flag in the metadata

4.7.2 Collections

a. Insert

- 1) The default option definition will be set if not specified.
- 2) Ingest will generate a unique collection ID with the following format: "C" + unique number + "-" + Data Center ID.
- 3) Beginning date/time shall be set to the least of: range date time, single date time, or periodic date time, whichever is specified in the metadata.
- 4) Ending date/time shall be set to the greatest of: range date time, single date time, or periodic date time, whichever is specified in the metadata
 - a) If periodic is used, the end time must be calculated using the periodic unit and cycle

b. Replace

- 1) All data in the collection will be replaced with the metadata of the replacement collection with the exception of
 - a) ECHO collection ID
 - b) Granule Spatial Representation
- 2) If the provider attempts to replace the Granule Spatial Representation, the collection will be rejected
- 3) If a granule references a platform in a collection, then the short names of referenced platforms, instruments, and sensors cannot be removed from the collection.

c. Delete

- 1) Deletes shall cascade to:
 - a) All items owned by the collection (as defined in the schema)
 - b) All granules in the collection

d. Partial Update/Add

- 1) Add/Updateable elements are:
 - a) Browse references
 - b) Visibility
 - c) Temporal
 - d) Restriction Flag
 - e) Delete Time
- 2) Delete-able elements are:
 - a) Temporal
 - b) Delete Time

- c) Restriction Flag

4.7.3 Granules

a. Insert

- 1) Beginning date/time shall be set to the least of: range date time or single date time, whichever is specified in the metadata.
- 2) Ending date/time shall be set to the greatest of: range date time or single date time, whichever is specified in the metadata
- 3) Ingest will generate a unique Echo granule ID with the following format: "G" + unique number + "-" + Data Center ID

b. Replace

- 1) All data in the granule will be replaced with the metadata of the replacement granule with the exception of
 - a) ECHO granule ID

c. Delete

- 1) Deletes shall cascade to
 - a) All items owned by the granule (as defined in the schema)
 - b) Browse links/references

d. Partial Update/Add

- 1) Add/Updateable elements are:
 - a) Browse references
 - b) Online access URL
 - c) Online resource URL
 - d) Measured Parameter/QA Flags (update only)
 - e) Additional attribute value
 - f) Visibility
 - g) Temporal
 - h) Day/Night Flag
 - i) Cloud Cover
 - j) Delete Time
 - k) Restriction Flag
- 2) Delete-able elements are:

- a) Online access URL
- b) Online resource URL
- c) Additional attribute value
- d) Temporal
- e) Day/Night Flag
- f) Cloud Cover
- g) Measured Parameter/QA Flag
- h) Delete Time
- i) Restriction Flag

4.7.4 Browse Images

- a. Browse Image links between collections and granules shall be sent in the collection and granule metadata
- b. Browse Metadata shall contain the following Browse Image Metadata information:
 - 1) Provider Browse ID
 - 2) InsertTime
 - 3) LastUpdate
 - 4) DeleteTime
 - 5) BrowseImageFileName
 - 6) Description
 - 7) FileSize
- c. If the association of the URL to it's item record fails (DB failure), then the Browse image is deleted from the Provider's repository (see Delete)
- d. **Insert**
 - 1) Browse image files shall be copied to the public browse path
 - 2) Browse image filenames shall be unique for a provider.
- e. **Replace**
 - 1) Browse image files shall be copied to the public browse path
 - 2) All data in the browse image will be replaced with the metadata of the replacement browse image

f. Delete

- 1) Deletes shall cascade to
 - a) All items owned by the Browse Image (as defined in the schema)
 - b) The browse image file
 - c) Browse links/references

g. Partial Add/Update

- 1) Partial adds/updates are not supported for browse

4.8. Ingest Reporting

4.8.1 Ingest Detail Report

- a. At the completion of each Ingest Job, Ingest will automatically generate an Ingest Detail Report for each ingest job with ingest job start/stop time information, data files received with the file sizes respectively, number of inserts/replaces/deletions/updates processed, number of rejections
 - 1) Ingest will save the Ingest Detail Report to an Archive
 - a) The Archive will be the Job Directory
 - 2) Ingest will put a copy of the Ingest Detail Report at the ftp site for Data Provider's access
- b. Ingest will not email the Ingest Detail Report. However, upon completion of the Ingest job, Ingest will email a job completion notification to both the Ingest Operator and the Data Provider.
 - 1) Emails to the Data Provider will be based on the email settings in the Data Provider's Configuration File
 - a) Whether to receive such emails
 - b) List of email addresses the notifications are to be sent to
- c. Reports in the Archive will be available to the Ingest Operators
- d. Data Providers will not have access to the Archive.
 - 1) For copies of reports no longer on the ftp site, Data Providers will contact the Ingest Operator
- e. The Ingest Detail Report will be in XML format
- f. For rejections, the Ingest Detail Report will list rejections along with the collections/granules/browses rejected.
- g. The report will record all alerts/errors resulting from Business Rule data validation errors
- h. The Ingest Detail Report is applicable to the following:
 - 1) Collections

- 2) Granules
- 3) Browse
- i. The Ingest Detail Report consists of two parts: overview and details
- j. The overview includes a summary of the items processed:
 - 1) Total Processed
 - 2) Total Inserted
 - 3) Total Replaced
 - 4) Total Updated
 - 5) Total Deleted
 - 6) Total Rejected
- k. The following details for each file will be included:
 - 1) File Name (this is the original file name as sent by the Data Provider)
 - 2) For each error code (rejection/alert), the following information will be included:
 - a) Rejection Reason
 - b) Item ID
 - (1) Dataset ID for collections
 - (2) Granule UR for granules
 - (3) Browse ID for Browse
 - (a) If there is no Provider Browser ID, the original XML metadata file name will be used

4.9. Provider-Specific Adapters

- a. If Cloud Cover is not specified in the metadata, the value shall be extracted from an additional attribute with a provider specified name
- b. If a granule spatial representation is not set then the provider's default will be used.
- c. Collection/Insert: The Visibility Flag will be set to the per-provider default collection visibility if not specified.
- d. Collection/Replace: The Visibility will be copied from the original item
- e. Granule/Insert: The Visibility Flag will be set to the per-provider default collection visibility if not specified.
- f. Granule/Replace: The Visibility will be copied from the original item

5. INGEST DATA ENTITIES

Ingest Data Entities are contained in two separate documents, *Ingest Data Entities* and *Ingest Configuration Entities*.

6. INGEST JOB STATES

The following lists the Ingest job states and the order they will occur on a job.

- a. RESOURCE_WAITING: Indicates the job is currently waiting for further resources to arrive from the provider.
- b. SEQUENCE_WAITING: Indicates the job is waiting on a sequence number earlier in the list to arrive and be processed.
- c. INPUT_ADAPTING: Indicates the job is having some of its files adapted from a provider specific format.
- d. SORTING: Indicates the job is having its files sorted by item type and action.
- e. LOAD_WAITING: Indicates the job is waiting for another job from the same provider to finish loading.
- f. LOADING: Indicates the job's metadata files are currently being validated and loaded into the database.
- g. REPORTING: Indicates a report is being generated for the job.
- h. CLEANING_UP: Indicates that cleanup is occurring on the job.
- i. COMPLETED: Indicates the job has been completed.

7. EMAIL NOTIFICATIONS CONTENT

The following sections define the content and formats to be used for e-mail notifications. Information within “<” and “>” indicates information that ingest will populate at runtime

7.1. Operational Emails

7.1.1 Start Ingest

Recipient:	Ingest Operator
CC:	
Reply to & Return Path:	
Subject:	<DATE> <TIME>, Ingest Start, <System>
Body:	Ingest on <System> started running at <DATE> <TIME>. <Ingest Web Services Port URL>
Signature:	

7.1.2 Stop Ingest

Recipient:	Ingest Operator
CC:	
Reply to & Return Path:	
Subject:	<DATE> <TIME>, Ingest Stopped, <System>
Body:	Ingest on <System> stopped running at <DATE> <TIME>.
Signature:	

7.2. Ingest Job-Specific Emails (per Data Provider)

7.2.1 No Input Received

Recipient:	Ingest Operator Data Provider Configured Email Addresses
CC:	
Reply to & Return Path:	
Subject:	Alert!! <DATE> <TIME>, No Input received from <PROVIDER>, <System>

Body:	This is a warning that no input files have been received from <PROVIDER> since <DATE> <TIME> on <System>. There may be a problem that requires further investigation.
Signature:	

7.2.2 Ingest Job Started

Recipient:	Ingest Operator Data Provider Configured Email Addresses
CC:	
Reply to & Return Path:	
Subject:	<DATE> <TIME>, <Provider> Ingest Job <JOB> Started <Job ID>, <System>
Body:	Ingest job <JOB> was started at <DATE> <TIME> on <System>
Signature:	

7.2.3 Ingest Job Completed

Recipient:	Ingest Operator Data Provider Configured Email Addresses
CC:	
Reply to & Return Path:	
Subject:	<DATE> <TIME>, <Provider> Ingest Job <JOB> Completed <Job ID>, <System> with the following state <Successfully Completed Deleted Completed with Errors>
Body:	Ingest job <JOB> <Job ID> for <Provider> was completed at <DATE> <TIME> on <System> Ingest Detail Report can be found at <ftp site> The Log Report can be found at <Log Report Path & File Name>
Signature:	

7.2.4 Out of Sequence Job

Recipient:	Ingest Operator
CC:	
Reply to & Return Path:	
Subject:	<DATE> <TIME>, <Provider> Job Out of Sequence, Expected Seq # <Expected Sequence #>, <System>
Body:	<DATE> <TIME>, <Provider> has an out of sequence job on <System>. The expected sequence number was <Expected Sequence #>
Signature:	

7.3. Ingest Errors/Failures

7.3.1 Error in Input Package and/or Table of Contents

Recipient:	Ingest Operator Data Provider Configured Email Addresses
CC:	
Reply to & Return Path:	
Subject:	<DATE> <TIME>, <Provider> Pkg or TOC Error <Job ID>, <System>
Body:	<DATE> <TIME>, <Provider> Error In The Input Package And/Or Table Of Contents <Job ID>, <System> <Description of error>
Signature:	

7.3.2 Fatal Error

Recipient:	Ingest Operator
CC:	
Reply to & Return Path:	
Subject:	Alert!! <DATE> <TIME>, Fatal Error occurred in Ingest <System>
Body:	The following fatal error occurred in <System> Ingest requiring its immediate shutdown at <DATE> <TIME>. Error: <EXCEPTION>
Signature:	

7.3.3 Error in Manifest File

Recipient:	Ingest Operator Data Provider Configured Email Addresses
------------	---

CC:	
Reply to & Return Path:	
Subject:	<DATE> <TIME>, <Provider> Manifest File Error <Job ID>, <System>
Body:	<DATE> <TIME>, <Provider> Error In the Manifest File <Job ID>, <System> <Description of error>
Signature:	

7.3.4 Job Requires Sequencing and Wait Time Exceeded

Recipient:	Ingest Operator Data Provider Configured Email Addresses
CC:	
Reply to & Return Path:	
Subject:	<DATE> <TIME>, <Provider> Out of Seq Timeout <Expected/Missing Sequence #>, <System>
Body:	<DATE> <TIME>, <Provider> on <System> has an out of sequence job and the expected sequence number was not received in <Out of Sequence Job Threshold Time>. The job has timed out Expected Seq #: <Expected/Missing Sequence #>
Signature:	

7.4. Pause/Resume Notifications

7.4.1 Input Detection Paused for Provider

Recipient:	Ingest Operator Data Provider Configured Email Addresses
CC:	
Reply to & Return Path:	

Subject:	<DATE> <TIME>, Input Detection for <Provider> has been Paused, <System>
Body:	<DATE> <TIME>, Input Detection for <Provider> has been Paused on <System>
Signature:	

7.4.2 Input Detection Resumed for Provider

Recipient:	Ingest Operator Data Provider Configured Email Addresses
CC:	
Reply to & Return Path:	
Subject:	<DATE> <TIME>, Input Detection for <Provider> has been Resumed, <System>
Body:	<DATE> <TIME>, Input Detection for <Provider> has been Resumed on <System>
Signature:	

7.4.3 Provider Paused

Recipient:	Ingest Operator Data Provider Configured Email Addresses
CC:	
Reply to & Return Path:	
Subject:	<DATE> <TIME>, <Provider> has been Paused, <System>
Body:	<DATE> <TIME>, <Provider> has been Paused on <System>
Signature:	

7.4.4 Provider Resumed

Recipient:	Ingest Operator Data Provider Configured Email Addresses
CC:	
Reply to & Return Path:	
Subject:	<DATE> <TIME>, <Provider> has been Resumed, <System>
Body:	<DATE> <TIME>, <Provider> has been Resumed on <System>
Signature:	

7.5. Report Notifications

7.5.1 Ingest Detail Report

Recipient:	Ingest Operator Data Provider
CC:	
Reply to & Return Path:	
Subject:	
Body:	
Signature:	

8. SECURITY

Security requirements and business rules are documented in "ECHO 10.0 Security Requirements" document.

9. VALIDS

- a. With ECHO 10, EDG Valid Processing will continue as is.
 - 1) ECS providers will send their valids file via email to ECHO Ingest Operations.
 - 2) Ingest Operations will run the valids file through a set of scripts
- b. ECHO 10 WIST valids will be retrieved using WIST's automated valids processing.

10. DESIGN CONSTRAINTS

- a. In cases where there is a conflict between Data Integrity and Performance, Data Integrity will take precedence.

11. NOTES

- a. Several "thresholds" have been identified in Ingest. These are briefly described in the Glossary (section 1) and described in detail in the "Ingest Configuration Entities" document.
- b. Editing the configuration file may require a third party editor such as XMLSpy or Microsoft Notepad.

12. ISSUES

12.1. General Issues

- a. Most providers do not specify a type with the granule additional attribute. NSIDC_ECS does but it's not clear what it means since its some random number.
- b. Issues with current Ingest Summary/Detail Reports that need to be addressed:
 - 1) LPDAAC would like a different format because they say the XML format is too difficult to parse
 - 2) Detailed report is too difficult to get. It's not automatically generated. Only OPS can generate it so Data Providers have to ask OPS and then OPS needs to find the time to do it.

Resolution: This has been addressed in ECHO 10. The report will be available to the Data Provider via ftp. If the Data Provider needs an old report that is no longer available on the ftp site, they will still have to notify Ingest Operations but Ingest Operations will not have to regenerate it, it will be in the Archive. [9/18/07]
 - 3) Too much unnecessary information (non-error info) in detailed report makes it difficult to find alerts/errors

Resolution: This has been addressed in ECHO 10. The report will contain detailed information for errors by filename and will only contain alert information. There will be a summary section at the beginning summarizing all activity—successful and unsuccessful. [9/18/07]

4) Manual process right now:

a) Ingest sends a list of files that could not be processed. Email contains attachments:

- (1) List of file names
- (2) Ingest Summary Report
- (3) Log entry (sometimes)

b) If report is too large to be sent via email the email fails and no further action is taken. This means OPS does not receive any information telling them the final status of the ingest job.

Resolution: This has been addressed in ECHO 10. The email will contain information as to the locations of the log and the Detail Report only. There is no longer an Ingest Summary Report [9/18/07]

c. Ingest Reporting: Original file name is essential. Will we still have the original file name readily available for reporting? Sorted file name is useless to Data Providers and OPS.

Resolution: Yes, the original file name will be included in the Ingest Detail Report [9/18/07]

12.1.1 File Handling

a. If the Ingest job is non-ECHO 10 compliant (i.e., using pre-ECHO 10 ingest format), then what do we do about the sequence # ?

Resolution: A sequence number of "-1" is used [9/18/07]